0

Research Report 1372

Cross-Validation of the Computerized Adaptive Screening Test (CAST)

Rebecca M. Pliske, Paul A. Gade, and Richard M. Johnson

Personnel Utilization Technical Area
Manpower and Personnel Research Laboratory

ari

U. S. Army
Research Institute for the Behavioral and Social Sciences

July 1984

Approved for public release, distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

	L. NEALE COSBY
EDGAR M. JOHNSON	**Colonel, IN**
Technical Director	**Commander**

Technical review by

Mary M. Weltin
Clessen J. Martin

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARI Research Report 1372	2. GOVT ACCESSION NO. AD-A163 148	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)		5. TYPE OF REPORT & PERIOD COVERED January 1983 - July 1984
		6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Rebecca M. Pliske, Paul A. Gade, and Richard M. Johnson		8. CONTRACT OR GRANT NUMBER(s) --
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333-5600		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263731A792
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences. 5001 Eisenhower Avenue Alexandria, Virginia 22333-5600		12. REPORT DATE July 1984
		13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computerized Adaptive Testing (CAT)
Computerized Adaptive Screening Test (CAST)
Enlistment Screening Test (EST), recruiting

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Computerized Adaptive Screening Test (CAST) was developed to provide an estimate at recruiting stations of prospects' Armed Forces Qualification Test (AFQT) scores. The CAST was designed to replace the paper-and-pencil Enlistment Screening Test (EST). The initial validation study of CAST indicated that CAST predicts AFQT at least as accurately as EST and that it is more efficient to use (Sands & Gade, 1983). This report summarizes the findings from a cross-validation study of CAST that used the following procedure. Prospects' CAST scores were recorded by recruiters in U.S. Army (Continued)

ARI Research Report 1372

20. (continued)

recruiting stations and were subsequently matched by social security number to applicant tapes from Military Entrance Processing Stations to obtain AFQT scores and relevant demographic data. These data were examined using regression, discriminant function, and cross-tabulation analyses to provide recruiters with information on how to interpret prospects' CAST scores. The correlation between CAST scores and AFQT scores was .80 for the cross-validation sample, whereas in the initial validation sample the correlation was .85. These data indicate that CAST scores are very good predictors of AFQT scores.

Research Report 1372

Cross-Validation of the Computerized Adaptive Screening Test (CAST)

Rebecca M. Pliske, Paul A. Gade, and Richard M. Johnson

Submitted by
Paul A. Gade, Chief
Personnel Utilization Technical Area

Approved as technically adequate
and submitted for publication by
Joyce L. Shields, Director
Manpower and Personnel
Research Laboratory

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

July 1984

Army Project Number 2Q263731A792	Joint Optical Information Network (JOIN)

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The Army faces a continuing and increasing demand to meet recruiting quantity and quality goals. Recent advances in computer technology and psychometric theory have made possible a new type of assessment technique, called computerized adaptive testing (CAT), that can provide accurate ability estimates based on relatively few test items. The Computerized Adaptive Screening Test (CAST) was designed to provide an estimate of a prospect's Armed Forces Qualification Test (AFQT) score at the recruiting station. Recruiters use prospects' CAST scores to determine whether the prospects should be sent to Military Entrance Processing Stations for further testing and to forecast the various options and benefits for which the prospects will subsequently qualify. This report will be used by the U.S. Army Recruiting Command (USAREC) to provide guidance to recruiters for the interpretation of CAST scores.

Edgar M Johnson
EDGAR M. JOHNSON
Technical Director

CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

EXECUTIVE SUMMARY

Requirement:

To cross-validate the Computerized Adaptive Screening Test (CAST) and to provide information that can be used by recruiters to predict prospective applicants' (prospects') Armed Forces Qualification Test (AFQT) scores from their CAST scores.

Procedure:

Prospects' CAST scores were recorded by recruiters in recruiting stations in the midwestern region of the United States. These scores were matched by social security number to applicant tapes from Military Entry Processing Stations (MEPSs) to obtain AFQT scores and relevant demographic data. These data were examined using regression, discriminant function, and cross-tabulation analyses. The results of these analyses were compared with the results of a previous validation study of CAST and a validation study of an alternative screening test called the Enlistment Screening Test (EST). An equal percentile equating of CAST scores and AFQT scores is summarized in a table that can be used by recruiters to interpret individual prospects' CAST scores.

Findings:

For the cross-validation sample, the correlation between CAST scores and AFQT scores was .80, whereas the correlation between CAST scores and AFQT scores in the previous sample was .85. The coefficient of determination ($\underline{r}^2$) for this sample was .63, as compared with a $\underline{R}^2$ value of .72 for the previous sample. A decrease in the amount of variance accounted for is to be expected, however, because the $\underline{R}^2$ value from an initial validation sample is always somewhat inflated as a result of capitalization on chance factors. The analyses of the data from the cross-validation sample indicate that CAST scores are good predictors of AFQT scores and that CAST is a reasonable alternative to EST. The correlation between EST scores and AFQT scores was estimated to be .83 ($\underline{r}^2$ = .69) in an initial validation sample that was composed of applicants from all the armed services. Cross-validation data on EST have never been reported. Because CAST is a computerized adaptive test it is considerably more efficient to use than EST.

Utilization of Findings:

This report will be used by the U.S. Army Recruiting Command (USAREC) to provide guidance to recruiters for the interpretation of CAST scores. It may also be used to make policy decisions regarding optimal cutpoints for CAST scores; however, the results reported should be interpreted with some degree of caution because they are based on a nonrandom sample of prospects from only one region of the United States.

CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

CONTENTS

Page

INTRODUCTION 1

Problem and Background 1
Computerized Adaptive Testing 2
Developing the CAST 3

CROSS-VALIDATION PROCEDURE 4

Description of CAST 4
Data Collection Procedure 4

RESULTS AND DISCUSSION 4

Regression Analyses 4
Comparison with Initial Validation Sample 7
Equipercentile Equating of CAST and AFQT 8
Probabilities of AFQT Classification 8
Comparison of CAST and EST 12

SUMMARY AND CONCLUSIONS 14

REFERENCES 17

LIST OF TABLES

Table 1. Sample demographics 6

2. Summary of regression analysis 7

3. Equipercentile calibration of CAST to AFQT 10

4. Probability estimates for AFQT category classification based on individual CAST scores 11

5. Percent of actual AFQT category classification given CAST scores for the applicants from the U.S. Army 4th Rctg Bde (MW) 13

6. Percent of actual AFQT category classification given EST scores for applicants from all services 13

Page

LIST OF FIGURES

Figure 1. Sample items from CAST 5

2. Frequency distributions for CAST and AFQT scores 9

CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

INTRODUCTION

The Computerized Adaptive Screening Test (CAST) was developed by the Navy Personnel Research and Development Center (NPRDC) and the Army Research Institute (ARI) to provide an estimate at recruiting stations of a prospective applicant's Armed Forces Qualification Test (AFQT) score. The CAST was designed to replace the paper-and-pencil Enlistment Screening Test (EST). The initial validation study of CAST indicated that CAST predicts AFQT at least as accurately as EST and that it is much more efficient to use (Sands and Gade, 1983). The research presented in this report summarizes the findings from a cross-validation study of CAST.

Problem and Background

All applicants for the armed services are given the Armed Services Vocational Aptitude Battery (ASVAB) to determine their eligibility for enlistment and their initial training assignment. AFQT is a linear composite of four ASVAB subtest scores: Word Knowledge (WK) and Paragraph Comprehension (PC) are combined to form an estimate of verbal ability that is combined with the Arithmetic Reasoning (AR) subtest score and one-half the Numerical Operations (NO) subtest score. The AFQT score is used by all services to determine an applicant's eligibility for enlistment. The ASVA3 is administered under very secure testing conditions either by the Department of Defense High School Testing Program or at a Military Entrance Processing Station (MEPS) or Mobile Examining Team (MET) site. Most prospective applicants are not tested in their high schools and must be sent to the MEPS/MET location for ASVAB testing, which entails transportation, food, and lodging expenses. Sending individuals who subsequently fail the ASVAB to the MEPS/MET locations is a waste of money and the recruiter's time; however, if prospective applicants who would have passed the ASVAB are not sent to the MEPS/MET locations for testing, the services lose valuable personnel.

The U.S. Army offers special options and skill training opportunities as enlistment incentives for qualified applicants. Special options include the Army College Fund, the 2-year Enlistment Option, and the Cash Bonus Enlistment Option. A qualified individual is a prospective applicant who has a high school diploma and scores at or above the 50th percentile on the ASVAB. If recruiters are to perform effectively for the Army, they need to know at an early stage of the interviewing process whether a prospective applicant is likely to qualify for enlistment incentive options. Failure to discuss options with prospective applicants who could have subsequently qualified for the options may result in lost sales contracts because the prospects remain ignorant of enlistment incentives that might have enticed them to join the Army. For example, Gade, Elig, Nogami, Hertzbach, Weltin, and Johnson (1984) showed that the majority of those who enlisted under the 2-year option said they would not have enlisted except for the 2-year option. Discussing options with prospective applicants who subsequently fail to qualify for the options can also result in lost contracts because these prospects are sold on features and benefits they cannot have, and they may fail to sign a contract at the MEPS.

Recruiters need to have an accurate prediction of AFQT scores at recruiting stations. They could use this information to determine which prospective applicants should be sent to the MEPS for additional testing. They could also use this information to tailor their sales presentation to discuss the features and benefits the Army has to offer applicants of different ability levels.

A paper-and-pencil test, called the Enlistment Screening Test (EST), is currently available for use by all the armed services at recruiting stations. Although EST scores provide accurate predictions of AFQT scores, EST has several drawbacks that are associated with most paper-and-pencil tests. The major drawbacks concern administrative errors and clerical burden (cf. Baker, Rafacz, and Sands, 1984). EST takes approximately 45 minutes to administer, and it must be hand-scored by the recruiter, which takes additional time and may introduce error. Because there are only two alternative EST forms, it is possible that prospective applicants might learn the items and eventually pass the test on repeated testing at different recruiting stations. All these problems can be eliminated because recent advances in computer technology and psychometric theory have made possible a new type of testing called computerized adaptive testing (CAT).

Computerized Adaptive Testing

An advance in psychometric theory, called Item Response Theory (Lord, 1980), has made it possible to adapt or tailor a test to the individual examinee. Unlike ability tests based on classical test theory, ability tests based on Item Response Theory (IRT) can provide comparable estimates of individuals' ability levels even when different individuals receive different sets of test items. In classical test theory all test parameters, such as item difficulty and discrimination indexes, are dependent on the specific test (i.e., a specific combination of items) and on the characteristics of the sample of individuals with whom the test was developed. In IRT, the focus is on test items and the probability of correct response to each item. The estimate of an individual's ability level is based on parameters associated with the specific items that individual received; these parameters are independent of the other items on the test and are also independent of the characteristics of the developmental sample. A detailed discussion of IRT is beyond the scope of this report. The interested reader is referred to Warm (1978) for an excellent introduction to IRT.

In traditional tests, each examinee responds to all items on the test. The traditional approach to test construction results in relatively poor measurement at the high and low ability extremes because many items on the test tend to be too difficult for the low-ability examinees or too easy for the high-ability examinees. In adaptive testing, each examinee receives the items that are appropriate to his or her ability level. The selection of each subsequent item is based on the examinee's previous response. If an examinee responded correctly to the last item, the next item will usually be more difficult than the previous one; but if the examinee's response to the last item was incorrect, the next item will usually be easier than the previous one. Adaptive testing makes it possible to construct tests that can discriminate equally well across all ability levels.

Although adaptive testing is possible without a computer, it is not very feasible because of the number of calculations and branching decisions that need to be made. In computerized adaptive testing, the computer presents each item and records the examinee's response. It computes an estimation of the examinee's ability level that determines the item that is administered next. A detailed discussion of the alternative procedures for making ability estimates and selecting subsequent items can be found in a report by McBride (1979).

In addition to improving the discriminability of tests, computerized adaptive tests are more efficient to use than traditional paper-and-pencil tests because they reduce testing time without sacrificing validity. Computerized adaptive tests also eliminate the need for manual scoring and recording, which can result in clerical errors, and they can provide immediate feedback on test results. Computerized adaptive tests reduce test compromise by eliminating test booklets that can be stolen and by administering different items to different individuals, making it more difficult for individuals to cheat. For all these reasons, a computerized adaptive test that can accurately predict a prospect's AFQT score at recruiting stations is a highly desirable recruiting tool.

Developing the CAST

The item pool for CAST was developed by researchers at the University of Minnesota (cf. Moreno, Wetzel, McBride, and Weiss, 1983) for use in the development of a computerized adaptive version of ASVAB (called CAT ASVAB). Initially there were three subtests developed, a Word Knowledge (WK) subtest, an Arithmetic Reasoning (AR) subtest, and a Paragraph Comprehension (PC) subtest. Moreno et al. provided a de facto pilot test of CAST in their research, which examined the relationship between corresponding ASVAB and CAT subtests. Thus, CAST was "pilot tested" with 270 male Marine recruits at the Marine Corps Recruit Depot in San Diego, Calif. The data from this pilot test indicated that the correlation between the optimally weighted CAST composite score and the AFQT score was .87. The data also indicated that the PC subtest did not improve the validity for predicting the AFQT score and that the PC items were extremely time consuming to administer. Therefore, the PC subtest was subsequently eliminated from CAST.

The initial validation study of CAST was conducted at the Los Angeles MEPS with a sample of 312 (251 male and 61 female) U.S. Army applicants (Sands and Gade, 1983). Each applicant received 20 WK items and 15 AR items from a pool of 78 WK items and 225 AR items. Sands and Gade analyzed the data collected at the Los Angeles MEPS to determine the optimal subtest length so that the predictive accuracy of CAST would be at least as high as that of EST (*r* = .83) with the shortest test administration time possible. Multiple correlation coefficients were computed for each of the 300 combinations of subtest length to develop the optimal prediction model for using CAST to forecast AFQT scores. Based on these analyses, a combination of 10 WK items and 5 AR items was recommended for the operational version of CAST. The correlation between this optimally weighted CAST score and actual AFQT score was .85. CAST is currently being implemented in Army recruiting stations throughout the United States.

There were two major limitations in the initial validation of CAST. First, the initial validation involved a relatively small sample (*N* = 312). Second, the test environment during the initial validation study was different from the test environment in which CAST is currently being used. In the initial validation, CAST was administered by a researcher at a MEPS to applicants who had already completed the ASVAB. The data in the research reported here were collected by recruiters at recruiting stations for prospective applicants before they were sent on to the MEPS for further processing.

CROSS-VALIDATION PROCEDURE

Description of CAST

CAST consists of 78 WK items and 225 AR items. All items are multiple choice items with a maximum of five response alternatives. The WK items generally deal with the definitions of words; the AR items generally deal with solving arithmetic work problems. Figure 1 illustrates the sample WK and AR items shown to subjects prior to testing. CAST uses the three-parameter logistic ogive item response model (Birmbaum, 1968); thus each test item has three parameters (discrimination, difficulty, and guessing) associated with it. Test items for CAST were chosen so that the discrimination parameter values would be greater than or equal to .78; the difficulty parameter values would range between +2 and -2; and the guessing parameter values would be less than or equal to .26. The ability estimate utilized in CAST is the Bayesian sequential scoring procedure discussed by Jensema (1977). The stopping rule is 10 WK and 5 AR items.

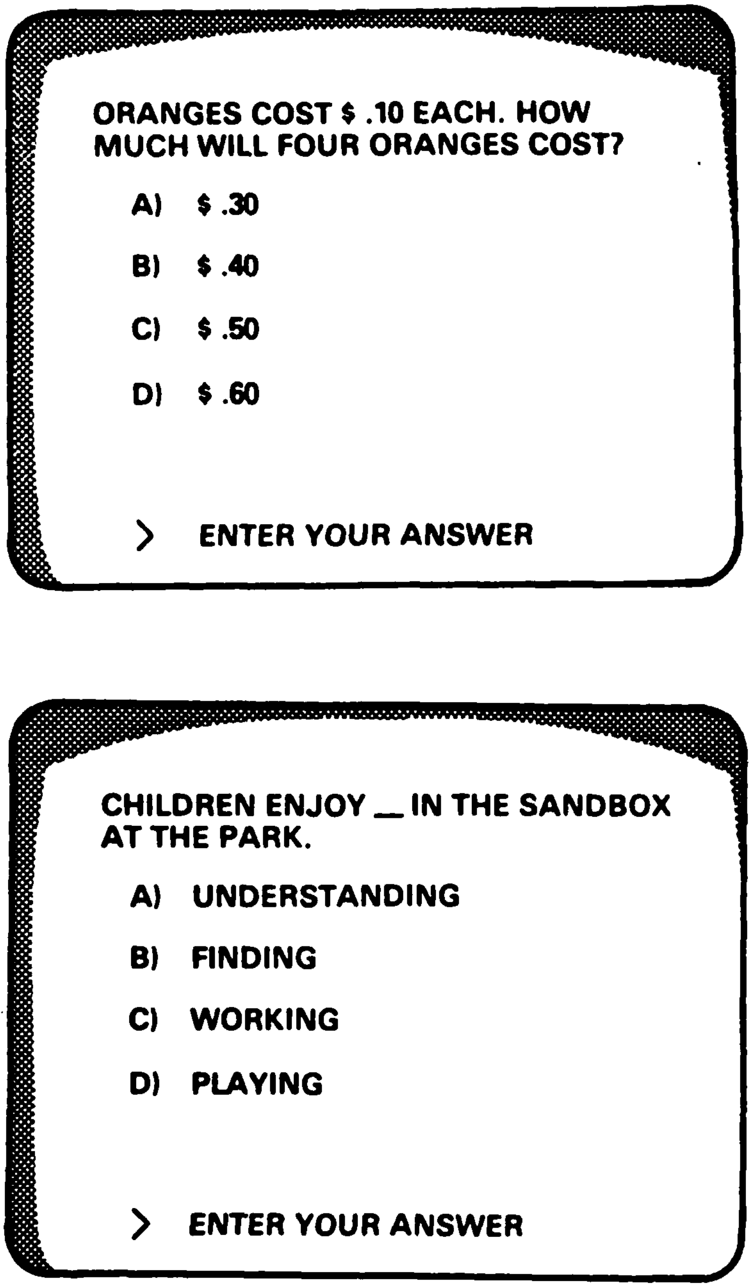
Data Collection Procedure

Prospects' CAST scores and social security numbers were recorded by recruiters in recruiting stations in the midwestern region of the United States during January and February, 1984. CAST is being introduced to recruiting stations by geographic region, and the midwestern region was the only fully operational region at the time of data collection. Recruiters were told to send all prospects for further testing, regardless of how poorly the prospects performed on CAST. The CAST scores and social security numbers of prospects were collected by the US Army Recruiting Command (USAREC) and forwarded to ARI for analysis. The CAST scores recorded by the recruiters were matched by social security number to applicant tapes from the MEPS to obtain AFQT scores and relevant demographic data on the applicants. Matching records were located for 1,962 applicants. The demographics of this sample are summarized in Table 1.

RESULTS AND DISCUSSION

Regression Analyses

The Pearson product moment coefficient calculated for CAST and AFQT scores in this sample is .80. This indicates that there is a strong, positive, linear relationship between CAST scores and AFQT scores. The coefficient of determination, $\underline{r}^2 = .63$, indicates that we can account for approximately 63% of the variability in applicants' AFQT scores by knowing their CAST scores. However, an $\underline{r}^2$ value of .63 also indicates that 37% of the variability in applicants' AFQT scores must be attributed to random error. Random factors which might influence the prediction of AFQT scores from CAST scores include anything that might



ORANGES COST $ .10 EACH. HOW MUCH WILL FOUR ORANGES COST?

A) $ .30

B) $ .40

C) $ .50

D) $ .60

> ENTER YOUR ANSWER

CHILDREN ENJOY __ IN THE SANDBOX AT THE PARK.

A) UNDERSTANDING

B) FINDING

C) WORKING

D) PLAYING

> ENTER YOUR ANSWER

Figure 1. Sample items from CAST.

Table 1

Sample Demographics

	Percentage
Gender	
Male	85
Female	15
Ethnic Group	
White	79
Nonwhite	21
Age	
16	2
17	21
18	21
19	18
20	10
21	7
22	5
23	4
24 or older	2
Prior Military Service	
No	93
Yes	7
Component	
RA	85
Reserves	15
Years of Education	
8	1
9	3
10	11
11	33
12	45
13	3
14	2
15	1
16	2

influence the prospect's performance on the test, such as test anxiety, physical fatigue, noisy test environment, etc.

There are other "nonrandom" factors that might influence the prediction of AFQT scores from CAST scores. These factors include demographic considerations such as the prospect's age, sex, and ethnic group. For example, CAST may be a better predictor of AFQT scores for white male prospects than for nonwhite female prospects. This would be an unfortunate finding because it would indicate that the test may be biased for certain subgroups of the population. In order to determine whether knowledge of certain demographic factors would affect the prediction of AFQT scores, we conducted a stepwise multiple regression analysis. The dependent measure in this analysis was the applicant's AFQT score, and the predictor variables were the applicant's CAST score, the six demographic variables listed in Table 1, and which alternative form of ASVAB the applicant took (e.g., Form 9A, 10X, etc.). Although all the alternate forms of the ASVAB used at the MEPS sites are parallel tests and should produce equivalent AFQT scores, it is possible that CAST scores may be better predictors of particular forms of ASVAB.

The results of this analysis, summarized in Table 2, indicate that the prospect's CAST score is the best predictor and, as reported previously, accounts for 63.3% of the variability in AFQT scores. Only two of the other predictors, number of years of education and ethnic group, accounted for any additional variance; and these two additional predictors increased the percentage of variance accounted for by only 0.5%. Therefore, it appears that having demographic information about prospects, in addition to their CAST scores, does not improve the ability to predict their AFQT scores.

Table 2

Summary of Regression Analysis

Variable entered	$\underline{R}^2$
Step 1 - CAST Score	.633
Step 2 - Years of Education	.637
Step 3 - Ethnic Group	.640

Comparison with Initial Validation Sample

For the cross-validation sample, the correlation between CAST scores and AFQT scores was .80, whereas the correlation between CAST scores and AFQT scores in the initial validation sample was .85. The coefficient of determination ($\underline{r}^2$) for this sample was .63, as compared with a $\underline{R}^2$ value of .72 for the initial validation sample. A decrease in the amount of variance accounted for ($\underline{R}^2$) is to be expected; the $\underline{R}^2$ value from an initial validation sample is always somewhat inflated because the procedures capitalize on chance relationships. The data from the cross-validation sample indicate that the operational version of CAST currently in use is a very good predictor of prospects' AFQT scores.

Equipercentile Equating of CAST and AFQT

There is a tendency to interpret a good predictor, such as CAST, as if it were a perfect predictor. However, it is incorrect to assume a prospect who scores 32 on CAST will subsequently score 32 on the ASVAB because the two tests have different scales. As mentioned in the introduction, AFQT scores are percentile scores based on a linear combination of four ASVAB subtest scores. Prospects' CAST scores are "raw scores" which are computed from an optimally weighted combination of WK and AR ability estimates. In order to make the scores more comparable for equating purposes, we converted the AFQT percentile scores to raw AFQT scores. When we plotted the frequency distribution of CAST and raw AFQT scores for the applicants in our sample (N = 1,962), we found that both sets of scores were approximately normally distributed as shown in Figure 2. The mean of the CAST scores is 49.67; the standard deviation is 18.35. The mean of the raw AFQT scores is 73.56; the standard deviation is 15.49.

Table 3 presents an equipercentile calibration of CAST with AFQT. This table was constructed by calculating the cumulative percent of scores which fell below each score in the frequency distributions of raw AFQT scores and CAST scores, and then equating the test scores based on these cumulative percentiles. Raw AFQT scores were then converted to AFQT percentile scores which is the typical form in which AFQT scores are presented. The AFQT percentile scores are based on the 1980 Youth Attitude norms. This process is illustrated in Figure 2 for the 25th, 50th, and 75th percentiles. For example, approximately 50% of the CAST scores fall below a CAST score of 50 and approximately 50% of the raw AFQT scores fall below the raw AFQT score of 77. Therefore, a CAST score of 50 is equivalent to a raw AFQT score of 77, which is equal to an AFQT percentile score of 49.

Probabilities of AFQT Classification

The equal percentile equating presented in Table 3 indicates the equivalent AFQT percentile score for a given CAST score. Recruiters should not, however, interpret the information presented in this table to mean that a prospect with a CAST score of 36 will always get an AFQT score of 29. If this were the case, then CAST might be the first perfect predictor test ever developed!

Actually, recruiters are not often concerned with the exact AFQT score the prospect subsequently receives. Instead, recruiters usually want to know into which mental test category (e.g., CAT IIIA, CAT IIIB, or CAT IV) a prospective applicant will subsequently be classified, because that is what determines whether the prospect will qualify for enlistment and for specific enlistment incentives. To help recruiters make this type of category prediction, we computed the probability of classification into the different mental categories based on prospects' CAST scores for our sample, and these results are shown in Table 4.

We used discriminant analysis to compute the "best-fitting" function that relates individual CAST scores to the four different AFQT categories and then, based on this function, computed the posterior probabilities of prospects being classified into the different AFQT categories based on his or her CAST score. To use the information in Table 4, locate the prospect's CAST score in the leftmost column and then, moving across the row, note the probabilities for each of

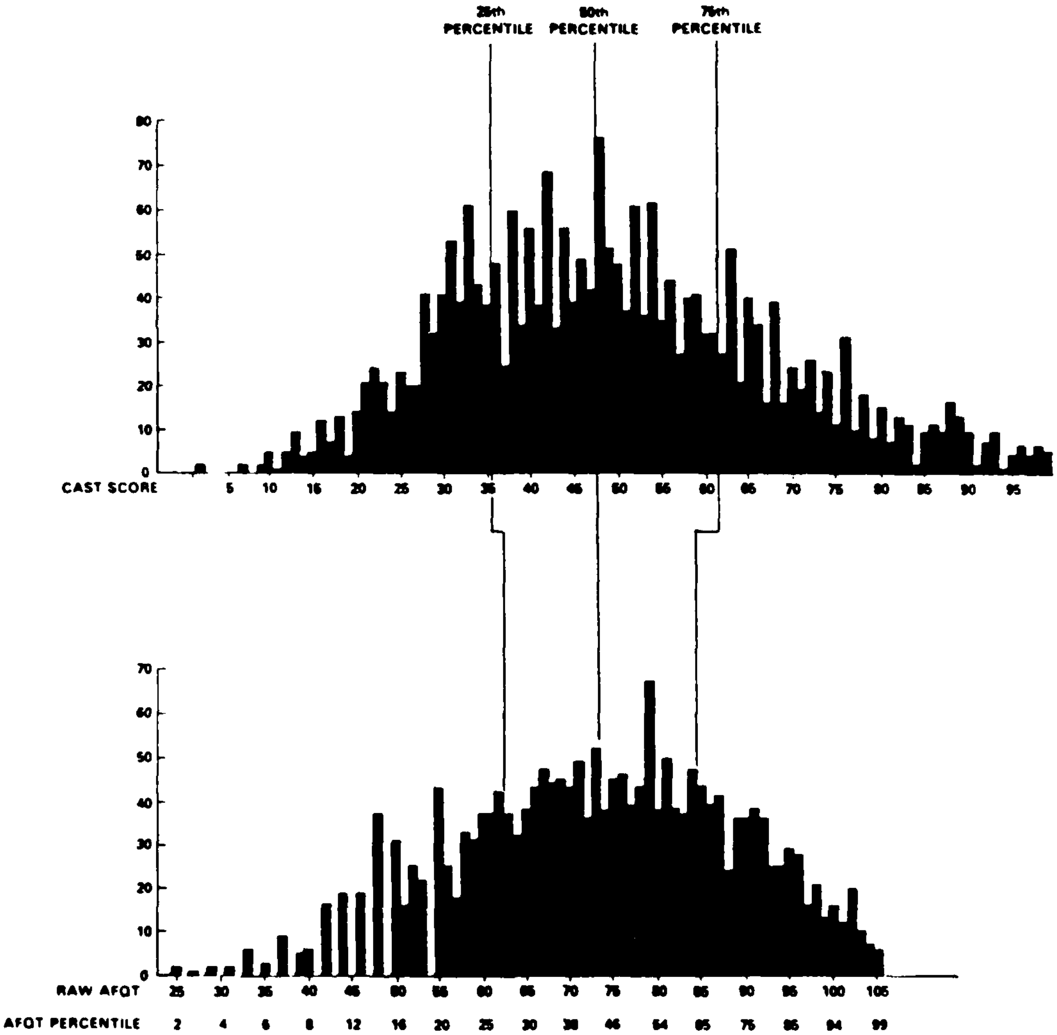


Figure 2. Frequency distributions for CAST and AFQT scores.

Table 3

Equipercentile Calibration of CAST to AFQT

CAST Score	Cumulative % below CAST Score	Raw AFQT Score	Cumulative % below Raw AFQT Score	AFQT Percentile Score
0-11	.5	33	.4	5
12-13	.6	37	.7	6
14-15	1.3	39	1.3	8
16-17	1.6	42	1.6	10
18-19	2.3	46	2.7	12
20-21	3.2	46	2.7	12
22-23	4.6	48	4.6	14
24-25	6.3	50	6.5	16
26-27	7.8	52	8.1	17
28-29	9.5	55	10.2	20
30-31	12.5	58	13.5	23
32-33	16.6	60	17.4	25
34-35	20.8	62	20.8	27
36-37	24.2	64	24.9	29
38-39	27.3	66	28.4	32
40-41	31.7	67	32.5	33
42-43	35.8	69	34.9	36
44-45	39.9	71	39.4	39
46-47	44.1	73	44.1	42
48-49	47.8	75	48.6	46
50-51	53.2	77	52.9	49
52-53	56.7	78	57.2	50
54-55	61.1	80	59.4	54
56-57	65.1	81	64.7	56
58-59	68.1	83	67.3	60
60-61	71.4	84	71.1	63
62-63	74.3	86	73.5	67
64-65	77.7	87	77.7	69
66-67	80.7	89	79.8	73
68-69	82.7	90	82.8	75
70-71	84.7	91	84.7	77
72-73	86.6	92	86.6	79
74-75	88.5	93	88.4	81
76-77	90.1	94	89.7	83
78-79	91.7	95	91.0	85
80-81	92.9	96	92.5	87
82-83	93.9	97	93.8	89
84-85	95.1	98	94.6	91
86-87	95.7	99	95.7	93
88-89	96.5	101	96.4	96
90-91	97.7	101	97.8	96
92-93	98.2	102	97.8	98
94-95	98.9	102	98.8	98
96-97	99.1	103	98.8	99
98-99	99.5	104	99.3	99

Table 4

Probability Estimates for AFQT Category Classification Based on Individual CAST Scores

	AFQT Category			
CAST	I/II	IIIA	IIIB	IV/V
Score	(65-100)	(50-64)	(31-49)	(0-30)
0-10	0	0	3	97
11-12	0	0	5	95
13-14	0	1	6	93
15-16	0	1	8	91
17-18	0	1	10	89
19-20	1	1	12	86
21-22	1	1	15	83
23-24	1	2	17	80
25-26	1	2	20	77
27-28	1	3	24	72
29-30	1	4	28	67
31-32	2	5	31	62
33-34	2	7	34	57
35-36	3	8	38	51
37-38	4	10	40	46
39-40	5	14	41	40
41-42	6	16	43	35
43-44	8	19	43	30
45-46	10	23	42	25
47-48	13	26	40	21
49-50	16	29	38	17
51-52	20	31	35	14
53-54	24	34	31	11
55-56	30	35	27	8
57-58	36	36	22	6
59-60	42	36	18	4
61-62	49	34	14	3
63-64	56	32	10	2
65-66	62	29	8	1
67-68	68	26	5	1
69-70	74	22	3	1
71-72	80	18	2	0
73-74	83	15	2	0
75-76	87	12	1	0
77-78	90	9	1	0
79-80	93	7	0	0
81-99	96	4	0	0

the four mental categories. For example, given a prospect with a CAST score of 47, there is a 13% chance for subsequent classification as a CAT I/II, a 26% chance as a CAT IIIA, a 40% chance as a CAT IIIB, a 21% chance as a CAT IV or below.

The horizontal lines in the table indicate important cutpoints between the AFQT categories. A prospect must have a CAST score of 37 or greater for the odds to be in favor of subsequent classification as a CAT IIIB or above; a prospect must have a CAST score of 51 or above for the odds to be in favor of subsequent classification as a CAT IIIA or above; and a prospect must have a CAST score of 63 or above for the odds to be in favor of subsequent classification as a CAT II or above.

Comparison of CAST and EST

CAST was developed to replace the paper-and-pencil Enlistment Screening Test (EST). Previous research (Sands and Gade, 1983) indicated that CAST predicted AFQT as least as well as EST and that it was much more efficient to use. The analyses of the data from the cross-validation sample indicate that CAST scores are good predictors of AFQT scores and that CAST is a reasonable alternative to the EST. The correlation between EST (Form 81A) scores and AFQT scores has been estimated to be .83 ($\underline{\ }^2$ = .689) in the initial validation sample which was composed of 486 applicants from all the armed services. Note that these values are very similar to those in our cross-validation of CAST. A cross-validation of EST has never been reported.

Tables 5 and 6 present data which indicate that both CAST and EST predict AFQT category classifications fairly well. Table 5 presents CAST scores and actual AFQT scores for the 1,962 applicants in our sample, classified according to the standard mental category cutpoints. Applicants were classified into the rows in Table 5 according to the cutpoints for CAST scores that are shown in Table 4. Applicants scoring above 63 on CAST were classified into the first row of the table labeled CAT I/II; applicants scoring between 51 and 62 were classified into the second row labeled CAT IIIA; applicants scoring between 37 and 50 were classified into the third row labeled CAT IIIB; and applicants scoring below 37 were classified into the last row of the table which is labeled CAT IV/V.

The columns in Table 5 represent the actual ASVAB AFQT category for each applicant. For example, 80% of the applicants who would have been classified as CAT Is and IIs based on their CAST scores were actually classified as CAT Is and IIs based on their ASVAB AFQT scores, but 13% were classified as CAT IIIAs based on their ASVAB AFQT scores, 5% were classified as CAT IIIBs, and 2% were classified as IVs or Vs. The data presented in row three of the table indicate that 43% of the applicants who would have been classified as CAT IIIBs based on their CAST scores were subsequently classified as CAT IIIBs based on their ASVAB AFQT scores; however, 28% were subsequently classified as CAT IIIAs and above, and 29% were subsequently classified as CAT IVs and below.

Table 6 presents EST scores and actual AFQT scores classified according to the standard mental category cutpoints. These data were adapted from a table in the Mathews and Ree (1982) paper. Unfortunately, the summary of the EST data provided by Mathews and Ree does not allow exact calculation of the category

Table 5

Percent of Actual AFQT Category Classification Given CAST Scores for the Applicants from the U.S. Army 4th Rctg Bde (MW)

CAST Category	AFQT Category I/II (65-100)	IIIA (50-64)	IIIB (31-49)	IV/V (0-30)
I/II (63-100)	80	13	5	2
IIIA (51-62)	33	40	21	6
IIIB (37-50)	8	20	43	29
IV/V (0-37)	1	4	25	70

Note. N = 1,962.

Table 6

Percent of Actual AFQT Category Classification Given EST Scores for Applicants from All Services

EST Category	AFQT Category I/II (65-100)	IIIA (50-64)	IIIB (31-49)	IV/V (0-30)
I/II (45-48)	86	13	1	0
IIIA (43-44)	50	37	11	2
IIIB (33-42)	16	28	38	18
IV/V (1-32)	2	6	23	69

Note. N = 869.

cutpoints, so we have had to approximate the EST category cutpoints. Applicants scoring above 45 on the EST were classified into the first row of the table labeled CAT I/II; applicants scoring 43 or 44 were classified into the second row labeled CAT IIIA; applicants scoring between 33 and 42 were classified into the third row labeled CAT IIIB; and applicants scoring below 33 were classified into the last row of the table which is labeled CAT IV/V.

The columns in Table 6 represent the actual ASVAB AFQT category for each applicant. For example, 86% of the applicants who would have been classified as CAT Is and IIs based on their EST scores were actually classified as CAT Is and IIs based on their ASVAB AFQT scores, but 13% were classified as CAT IIIAs based on their ASVAB AFQT scores, and 1% were classified as CAT IIIBs. The data presented in row three of the table indicates that 38% of the applicants who would have been classified as CAT IIIBs based on their EST scores were subsequently classified as CAT IIIBs based on their ASVAB AFQT scores; however, 44% were subsequently classified as CAT IIIAs and above, and 18% were subsequently classified as CAT IVs and below. The data presented in Tables 5 and 6 indicate that both CAST and EST are good predictors of prospective applicants' subsequent classification into AFQT categories.

SUMMARY AND CONCLUSIONS

The Computerized Adaptive Screening Test (CAST) was developed to provide an estimate of a prospect's Armed Forces Qualification Test (AFQT) score on the Armed Services Vocational Aptitude Battery (ASVAB). The CAST was developed to replace the paper-and-pencil Enlistment Screening Test (EST). The data collected in the initial validation study of CAST (Sands and Gade, 1983) and in the cross-validation reported here indicate that CAST predicts AFQT scores at least as accurately as the EST; and because CAST is a computerized adaptive test, it is more efficient to use.

Although CAST as it is currently operationalized in the field is a very good predictor of AFQT scores, it could be modified so that it could make even more accurate predictions. As discussed previously, CAST is based on Item Response Theory, an advance in psychometric theory that permits item parameters to be calculated for each individual test item. The test items selected for inclusion in the operational version of CAST were chosen so that CAST would discriminate equally well across all ability levels. CAST was also designed to provide a point estimate of a prospect's AFQT score (e.g., 86, 71, 35) rather than a category estimate (e.g., CAT IIIA, CAT IIIB). CAST might be of greater use to recruiters if it were modified to provide a more accurate estimate of AFQT category classification.

Three changes could be made in the operational version of CAST to improve the accuracy with which it predicts AFQT scores at the critical cutpoints for AFQT category classifications. First, the optimal weighting of the CAST WK and AR subtests for predicting AFQT scores was determined for making point estimates (Sands and Gade, 1983), and it is not necessarily (nor likely) the optimal weighting for predicting AFQT categories. Individual item data need to be collected from a large sample of prospects from recruiting stations across the country. Discriminant function analyses of these data could specify the optimal weighting of the CAST subtests for predicting subsequent AFQT classification. Second, new test items could be developed for CAST that would have very high discriminability parameters for the critical cutpoints for AFQT category classification. The development of new items would improve the accuracy with which CAST could discriminate between individuals who would subsequently be classified as CAT IIIAs and CAT IIIBs, or between individuals who would subsequently be classified as CAT IIIBs or CAT IVs. Third, a new item selection procedure could be implemented that would be specifically designed to optimize the prediction of AFQT category classifications. Although the Bayesian sequential scoring procedure currently used in the operational version of CAST is an appropriate ability estimation procedure for the prediction of the continuous AFQT scale, it may not be the most appropriate procedure for prediction of AFQT categories. Alternative procedures need to be developed and tested.

Because CAST is a computerized adaptive test, individual item data can be collected via the computer while the prospect is taking an operational version of the test. Therefore, the collection of the data that is necessary for the future refinement and improvement of CAST can be totally "invisible" to the prospect taking CAST. In addition to responding to the ten WK items and five AR items currrently used tc estimate ASVAB AFQT, prospects in selected test stations could be administered several additional test items. In this way, we could collect item "calibration" data on very large samples of prospects so that new test items could be developed. The operational version of the CAST software is currently being modified to record the individual item data necessary for future improvements to CAST. The collection of item response data should make it possible for CAST to be continually modified to meet the current needs of the U.S. Army Recruiting Command.

REFERENCES

Baker, H.G., Rafacz, B.A. and Sands, W.A. (1984). *Computerized Adaptive Screening Test (CAST): Development of use in military recruiting stations* (NPRDC Report No. 84-17). San Diego, CA: Navy Personnel Research and Development Center.

Birmbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds) *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Gade, P.A., Elig, T.W., Nogami, G.Y., Hertzbach, A., Weltin, M., and Johnson, R.M. (1984). *Motives, incentives, and key influencers for enlistment, reenlistment, and attrition in the US Army*. U.S. Army Research Institute for the Behavioral and Social Sciences. Paper presented at the NATO Symposium on Motivation and Morale, Brussels, Belgium.

Jensema, C.G. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement*, *1*, 111-120.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mathews, J.J. and Ree, M.J. (1982). *Enlistment Screening Test Forms 81a and 81b: Development and Calibration* (AFHRL Report No. 81-54). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

McBride, J.R. (1979). *Adaptive mental testing: The state of the art* (ARI Report No. 423). Alexandria, VA: U.S. Army Research Institute. (NTIS No. ADA088000).

Moreno, K.E., Wetzel, C.D., McBride, J.R., and Weiss, D.J. (1983). *Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests* (NPRDC Report No. 83-27). San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. ADA131683)

Sands, W.A., and Gade, P.A. (1983). An application of computerized adaptive testing in U.S. Army recruiting. *Journal of Computer-based Instruction*, *10*, 87-89.

Warm, T.A. (1978). *A primer of item response theory*. (USCGI Report No. 941278). Oklahoma City, OK: U.S. Coast Guard Institute.